Semiquantitative Group Testing in at Most Two Rounds

Mahdi Cheraghchi
Department of EECS
University of Michigan, Ann Arbor, MI
Email: mahdich@umich.edu

Ryan Gabrys
Naval Information Warfare Center,
San Diego, and UIUC, Urbana, IL
Email: ryan.gabrys@gmail.com

Olgica Milenkovic
Department of ECE
University of Illinois, Urbana, IL
Email: milenkov@illinois.edu

***Abstract*—Semiquantitative group testing (SQGT) is a pooling method in which the test outcomes represent bounded intervals for the number of defectives. Alternatively, it may be viewed as an adder channel with quantized outputs. SQGT represents a natural choice for Covid-19 group testing as it allows for a straightforward interpretation of the cycle threshold values produced by polymerase chain reactions (PCR). Prior work on SQGT did not address the need for adaptive testing with a small number of rounds as required in practice. We propose conceptually simple methods for 2-round and nonadaptive SQGT that significantly improve upon existing schemes by using ideas on nonbinary measurement matrices based on expander graphs and list-disjunct matrices.**

I. Introduction

Group testing (GT) is a scheme designed to efficiently identify a small set of subjects with a particular property (standardly referred to as defectives) within a large population, first introduced by Dorfman [1] and further studied in many other works, including [2]–[4]. Group testing entails testing a collection of carefully selected subpopulations and reporting for each subgroup a binary answer: A positive answer is indicative of the existence of at least one defective in the subgroup while a negative answer implies the absence of defectives. Given that screening protocols are extensively used in engineering and science, group testing has found widespread applications in communication theory, signal processing, computer science, and computational biology [3], [5].

Many different variants of group testing have been proposed in the literature [1], [3], [6]. These include threshold group testing proposed by Damaschke [7] and quantitative (additive) group testing studied by Lindström and Du and Hwang [6], [8], [9]. In the latter case, the test results report the exact number of defectives in the test subpool. In the former case, if the number of defectives in a test is smaller than a lower threshold, the test outcome is negative; if the number of defectives is larger than an upper threshold, the test outcome is positive; otherwise, the result is arbitrary (positive or negative). To bridge the two above described paradigms, Emad and Milenkovic [10]–[12] introduced the notion of semiquantitative group testing (SQGT). SQGT represents a unifying framework of a number of testing protocols, including conventional, quantitative and gapless threshold group testing and the schemes by D'yachkov and Rykov [13], [14]. In SQGT, the result of a test is a nonbinary value that depends on the number of defectives through a fixed set of thresholds. The SQGT model may also be viewed as a quantitative group testing method followed by a quantizer. The original motivation for introducing SQGT models is genotyping; more recently, the model has been used by Gabrys et al. [15] to describe the test outcomes of a Covid-19 testing process known as real-time reverse-transcriptase polymerase chain reaction (PCR).

In *nonadaptive* SQGT, each subject is assigned a unique binary or nonbinary indicator word of length equal to the total number of tests. These indicators are arranged column-wise in a *test matrix*. Each coordinate in the codeword assigned to a subject corresponds to a test, and its value reflects the "concentration" of the sample corresponding to the given subject in the test. Note that the concentrations are nonnegative integers that usually correspond to the number of units of the genetic material of an individual subject. Two families of nonadaptive SQGT codes, SQ-disjunct and SQ-separable, were analyzed in [11], [12]. In the same work, a number of constructions for nonadaptive uniform and nonuniform (quantized) SQGT codes were presented but no results were reported for adaptive tests. The more recent work [15] introduced the first combinatorial and probabilistic adaptive SQGT (ASQGT) schemes, the former extending the work of Hwang [16] on generalized binary group testing. The proposed combinatorial ASQGT schemes involve what is referred to as parallel and deep search methods that lead to a relatively large number of testing rounds. This is an undesirable feature for practical implementations of SQGT in Covid-19 testing.

Here, we describe the first known combinatorial two-round adaptive SQGT (ASQGT) for a special selection of (quantization) thresholds studied in [15]. The scheme uses $O\left(\frac{d \log\log \tau}{\log \tau} \log \frac{n}{d}\right)$ tests for n subjects, d defectives and τ SQGT thresholds. It builds upon the ideas of list-disjunct group testing [17] and like the approach [15] uses nonbinary test matrices obtained by careful linear combining of the rows of a binary disjunct matrix. The described two-round ASQGT protocol differs from the information-theoretic bound only by about a factor $\log \tau$. We then proceed to improve existing nonadaptive protocols by extending the construction of Porat and Rothschild [18].

M. Cheraghchi's research was partially supported by the National Science Foundation under Grant No. CCF-2006455.

The paper is organized as follows. Sections III describes our main result, the first known two-round ASQGT. Section IV presents new nonadaptive SQGT schemes that significantly improve upon previous constructions [11], [12] and imply new upper bounds for nonadaptive SQGT.

II. Terminology, GT Background, and Bounds

We start with some relevant terminology. All parameters are denoted by small-case letters, while vectors and matrices are denoted by bold-face small-case and capitalized Latin letters, respectively. Entries of the vectors are indexed by subscripts while matrix entries are indexed by pairs of integers within parentheses. Unless stated otherwise, all logs are to base 2.

Assume that there are $n > 1$ test subjects labeled by elements in $[n] := \{1, \ldots, n\}$ among which $d < n$ are defective (i.e, infected). In conventional group testing, we summarize the set of tests through a binary matrix $\mathbf{B}^{m\times n}$ in which every column of the matrix uniquely characterizes an individual and each row represents a test. The $(i,j)^{\text{th}}$ entry of $\mathbf{B}$, $\mathbf{B}(i,j)$, equals 1 if and only if the individual labeled j is included in the i^{th} test. Let $\mathbf{t}_I \in \{0,1\}^m$ denote the binary vector that results from m tests using $\mathbf{B}$, assuming that the set of infected individuals equals $I \subset [n]$, with $|I| \leqslant d$. Whenever clear from the context, we omit the subscript I. In conventional group testing $\mathbf{t}_I(l) = 1$ if and only if the l^{th} test includes at least one element from I. Let $\mathbf{t}_L \in \{0,1\}^m$ be defined analogously for another set $L \subset [n]$. We say that a set L is consistent with I if $\mathbf{t}_L \leqslant \mathbf{t}_I$ entrywise.

The matrix $\mathbf{B}^{m\times n}$ is termed d-disjunct if no vector $\mathbf{t}_I$ for $|I| \leqslant d$ contains in its support a column of $\mathbf{B}$ not indexed by I. The disjunctness property ensures that the test results obtained from B uniquely identify the set of defectives. A matrix $\mathbf{B}$ is termed (d,ℓ)-list-disjunct if the tests output a superset of the defectives of size at most $\ell + d$; for such a matrix, the size of any list L consistent with I is at most $\ell + d$. Clearly, a matrix $\mathbf{B}$ which is d-disjunct is equivalent to one which is $(d,0)$-list-disjunct. The notion of list-disjunct matrices was explicitly formulated (in an equivalent form) in [19] and is also essentially equivalent to what was defined earlier in [20].

We review the following known results pertaining to the existence of (d,ℓ)-list-disjunct test matrices $\mathbf{B} \in \{0,1\}^{m\times n}$ with $\ell = \mathcal{O}(d)$ and $m = \mathcal{O}(d \log \frac{n}{d})$. First, note that it is straightforward to see that for a maximal L one has $I \subseteq L$. Therefore, as noted in [21], the existence of a (d,ℓ)-list-disjunct test matrix $\mathbf{B} \in \{0,1\}^{m\times n}$ with $\ell = \mathcal{O}(d)$ naturally implies a two-round testing scheme: The first round of tests is governed by the rows of $\mathbf{B}$ while the second round involves individually testing subjects in L. Randomized and explicit constructions of list-disjunct matrices exist, particularly via expander graphs [14], [17], [19]–[21]. The best known construction which achieves an optimal number of rows and nearly linear time recovery (in the number of rows) is given by [22].

The best lower bound on the number of tests necessary for an adaptive ASQGT scheme was established in [15] via a simple counting argument and the bound equals $\frac{d}{\log\tau}\log\frac{n}{d}$. In the next section, we establish the existence of a two-round scheme that differs from this lower bound by a factor of $\log\log\tau$ only. For the single-round setting, using a variation of the argument employed by Füredi [23] in the context of cover-free codes, one can show that the corresponding number of tests scales as $\frac{d^2}{(\log\tau)^3}\log_d n$ whereas the construction from Section IV implies the existence of a scheme that requires at most $\frac{d^2}{\log\tau}\log n$ tests. This lower bound applies to not only general nonadaptive SQGT, but in fact the particular *saturation model* as well, which is the focus of this work. The derivation of the bound is relegated to the full version of the paper.

III. Two-Round ASQGT

Let $\mathcal{G}$ be a bipartite graph with a vertex partition $\mathcal{P}$ (people) and $\mathcal{T}$ (tests) such that every vertex in $\mathcal{P}$ has degree k (i.e., k neighbors) and $|\mathcal{P}| = n$, $|\mathcal{T}| = m$. We say that $\mathcal{G}$ is an (α,β)-expander if every $P \subseteq \mathcal{P}$ of size at most $\alpha|\mathcal{P}|$ has at least $\beta|P|$ neighbors in $\mathcal{T}$. The values of the parameters n, m are dictated by the expansion factors α, β. It is also worth pointing out that *explicit* constructions of expander graphs with parameters of interest in our derivations may not be known, but their existence is guaranteed via probabilistic arguments. We say that a set of vertices $T \in \mathcal{T}$ is covered by a set $P \subseteq \mathcal{P}$ if for every vertex $t \in T$, there exists a vertex $p \in P$ which is connected to t. We say that a vertex $t \in \mathcal{T}$ is uniquely covered (or a unique neighbor) of P if it is the neighbor of exactly one vertex $p \in P$. Henceforth, for a set of vertices P, let $N(P) \subseteq \mathcal{T}$ denote the neighbors of P and let $N_u(P) \subseteq \mathcal{T}$ denote the set of unique neighbors of P. Furthermore, we say that a vertex $t \in \mathcal{T}$ is covered h times by P if it is connected to exactly h different vertices in P. The next results may be obtained through a straightforward modification of existing results.

Lemma 1. *[19] Suppose that $\mathcal{G}$ is an (α,β)-expander where every vertex in $\mathcal{P}$ has k neighbors and $\beta > \frac{3k}{4}$. Let $I \subseteq \mathcal{P}$ be a subset of size at most $|I| \leqslant d$. Then for any $P \subseteq \mathcal{P}$ such that $P \cap I = \varnothing$, $|P| \geqslant |I| + 2$ and $|P \cup I| \leqslant \alpha|\mathcal{P}|$, we have:*

$$\left|N_u(P \cup I) \backslash N(I)\right| \geqslant k.$$

Thus, given the previous lemma, it follows that there exists at least one test in $N(P \cup I)$ that is not covered by an element of I. Using this observation, we construct the $m \times n$ binary matrix $\mathbf{B}$ as follows. Suppose that $\mathcal{G}$ is an expander as previously described. We assume that the vertices in $\mathcal{P}$ and $\mathcal{T}$ are lexicographically ordered so that we can refer to the i^{th} vertex in $\mathcal{T}$ as i and the j^{th} vertex in $\mathcal{P}$ as j. Then, for $i \in \mathcal{T}$ and $j \in \mathcal{P}$,

$$\mathbf{B}(i,j) = \begin{cases} 1, & \text{if an edge exists between } i \text{ and } j \text{ in } \mathcal{G}, \\ 0, & \text{otherwise.} \end{cases} \tag{1}$$

Thus, as a result of the construction for $\mathbf{B}$, we see that we can uniquely associate each column of $\mathbf{B}$ with a vertex in $\mathcal{P}$ and each row of $\mathbf{B}$ with a vertex in $\mathcal{T}$.

The next two results follow immediately from the previous discussion.

Corollary 2. *Suppose we are given two sets $I, L \subseteq \mathcal{P}$ such that $I \subseteq L$. If L is consistent with I under $\mathbf{B}$ and $|L| \leqslant \alpha|\mathcal{P}|$, then*

$$|L| < 2|I| + 2.$$

Lemma 3. *Suppose $\mathbf{B}$ is as defined in (1) and the set of infected individuals satisfies $|I| \leqslant d$. Then, testing with $\mathbf{B}$ recovers a set $L \subseteq \mathcal{P}$ such that $|L| = \mathcal{O}(d)$ and $I \subseteq L$.*

The following lemma is known [3] and follows from a standard randomized construction:

Lemma 4. *Suppose that $\alpha = \frac{2d+2}{n}$ and let $m = 8e^2 k\alpha n$, where e is the base of the natural logarithm. Then, for $k = \mathcal{O}(\log \frac{1}{\alpha})$ there exists an (α, β)-expander graph $\mathcal{G}$ with bipartition $\mathcal{P}, \mathcal{T}$ such that $|\mathcal{P}| = n$ and $|\mathcal{T}| = m$, and $\beta = \frac{3k}{4}$.*

The previous result implies the following theorem.

Theorem 5. *There exists a conventional two-stage GT scheme that requires at most*

$$\mathcal{O}\left(8e^2 d \log \frac{n}{d}\right)$$

tests and can identify a set of infected individuals of size at most d from a population of size n.

We remark that the best known explicit constructions of bipartite expanders are still inferior to the optimal bounds achieved by random expanders in Lemma 4. For example, using [24] one can get $O(d^{1+\alpha}(\log n)^{O(1/\alpha)})$ tests for any fixed $\alpha > 0$, and [25] would achieve $O(d\exp((\log\log n)^3))$ tests, similar to the derivation in [20].

We now discuss how to use the matrix $\mathbf{B}$ to design a specialized two-round SQGT testing scheme for $\tau > 2$.

We focus on a special case of uniform SQGT *with saturation* [15] for which we are given τ thresholds. The test outcome vector for a set I of defectives is such that $\mathbf{s}_I(l) = 0$ if the l^{th} test includes no defectives, $\mathbf{s}_I(l) = 1$ if the l^{th} test includes 1 defective, $\ldots$, $\mathbf{s}_I(l) = \tau - 2$ if the l^{th} test includes $\tau - 2$ defectives and $\mathbf{s}_I(l) = \tau - 1$ if the number of defectives in the l^{th} test exceeds $\tau - 2$. To simplify the notation, we assume that $\tau = (4\gamma)^\gamma$, for some positive integer γ.

We show the existence of a two-round testing scheme that differs from the information theoretic lower bound from [15] by only a factor of roughly $\log\tau$. As discussed earlier, we only focus on the first round, since the second one is straightforward. The key idea used to construct the test matrix for the first round is to start with list-disjunct expander-based binary test matrix and then merge the rows via specialized linear combinations to reduce the number of tests and increase the size of the alphabet used for the codebook.

We start by introducing two matrices $\mathbf{S}^{(1)}$ and $\mathbf{S}^{(2)}$ that will be subsequently concatenated into the "global" SQGT matrix $\mathbf{S} = \begin{bmatrix}\mathbf{S^{(1)}} \\ \mathbf{S^{(2)}}\end{bmatrix}$. Let $\mathbf{B}$ be as defined in (1) and for simplicity, assume that $\gamma \mid m$. Then, for $i \in [1, \frac{m}{\gamma}]$ and $j \in [1, n]$, we set

$$\begin{aligned}\mathbf{S}^{(1)}(i,j) &= \mathbf{B}((i-1)\gamma+1, j) + (4\gamma)\mathbf{B}((i-1)\gamma+2, j) \quad (2)\\ &\quad + (4\gamma)^2\mathbf{B}((i-1)\gamma+3, j) + \cdots + (4\gamma)^{\gamma-1}\mathbf{B}(i\gamma, j);\\ \mathbf{S}^{(2)}(i,j) &= \mathbf{B}((i-1)\gamma+1, j) + \mathbf{B}((i-1)\gamma+2, j) \quad (3)\\ &\quad + \mathbf{B}((i-1)\gamma+3, j) + \cdots + \mathbf{B}(i\gamma, j).\end{aligned}$$

Note that both $\mathbf{S}^{(1)}$ and $\mathbf{S}^{(2)}$ are obtained linear combination of rows of $\mathbf{B}$, but the scaling factors are different. The SQGT test matrix $\mathbf{S}$ has $2\frac{m}{\gamma}$ rows and consequently the same number of tests. The tests involve taking an integer number of sample units dictated by the nonbinary entries in $\mathbf{S}$. The nonbinary (semi-quantitative) test outcome vector will be denoted by $\mathbf{s}$.

Let $E(a)$ denote the (4γ)-ary expansion of the natural number a in vector form. More precisely, if $a = a_0 + a_1 4\gamma + a_2 (4\gamma)^2 + \cdots + a_{\gamma-1}(4\gamma)^{\gamma-1}$, then $E(a) = \left(a_0, a_1, \ldots, a_{\gamma-1}\right)$, where $a_i, i \in [0, 4\gamma-1]$. Our decoding procedure operates as follows. Suppose that $\mathbf{s}^{(1)} = (s_1^{(1)}, \ldots, s_{\frac{m}{\gamma}}^{(1)})$ represents the results of the (quantized) testing using the matrix (2). We apply the map E to $\mathbf{s}^{(1)}$ entrywise. We then use an expander-based decoding procedure on this vector to recover a "noisy" set of test values - the "noise" is due to the that the matrix $\mathbf{S}^{(2)}(i,j)$ can handle only up to 4γ defectives.

To this end, let $\mathbf{s}' = \left(E(s_1^{(1)}), E(s_2^{(1)}), \ldots, E(s_{\frac{m}{\gamma}}^{(1)})\right) = (s_1', s_2', \ldots, s_m')$ and let $\hat{\mathbf{t}}^{(b)} = \left(\lceil\frac{s_1'}{\tau}\rceil, \ldots, \lceil\frac{s_m'}{\tau}\rceil\right) = \left(\hat{t}_1^{(b)}, \hat{t}_2^{(b)}, \ldots, \hat{t}_m^{(b)}\right) \in \{0,1\}^m$. Note that $\hat{\mathbf{t}}_i^{(b)} = 0$ if $s_i' > 0$ and zero otherwise. For shorthand, we write $f_{\tau\to b}\left(\mathbf{s}^{(1)}\right) = \hat{\mathbf{t}}^{(b)}$. We have the following claim.

Claim 6. *Let $\mathbf{t} \in \{0,1\}^m$ denote the test output based on the binary matrix $\mathbf{B}$, let $\mathbf{s}^{(1)}$ be the test output generated via $\mathbf{S}^{(1)}$ and let $\hat{\mathbf{t}}^{(b)}$ be as defined above.Then,*

$$d_H\left(\hat{\mathbf{t}}^{(b)}, \mathbf{t}\right) \leqslant \frac{dk}{4}.$$

Proof: Let $f_{\tau\to b}(s_i^{(1)}) = (t_{\gamma(i-1)+1}, \ldots, t_{\gamma i})$ be the mapping that corresponds to $\mathbf{s}_i^{(1)}$. For some $j \in [0, \gamma-1]$, let vertex $(i-1)\gamma + j \in \mathcal{T}$ be covered $\geqslant 4\gamma$ times. Such a vertex may be in error (due to the use of the 4γ-ary expansion). Since the set $I \subseteq \mathcal{P}$ has at most $|I|k$ neighbors in $\mathcal{T} \subseteq \mathcal{G}$, it follows from an averaging argument that

$$\left|\left\{(i,j) : \text{vertex } (i-1)\gamma + j \in \mathcal{T} \text{ is covered} \geqslant 4\gamma \text{ times}\right\}\right| \leqslant \frac{|I|k}{4\gamma}.$$

Let $(i-1)\gamma + \ell \in \mathcal{T}$ be a vertex in $\mathcal{T}$ which is covered at least 4γ times (if no such vertex exists, we are error-free and do not have to prove anything further). In this case we may have $f_{\tau\to b}\left(s_{(i-1)\gamma+\ell}^{(1)}\right) = \left(\hat{t}_{(i-1)\gamma+1}^{(b)}, \hat{t}_{(i-1)\gamma+2}^{(b)}, \ldots, \hat{t}_{\gamma}^{(b)}\right) \neq \left(t_{(i-1)\gamma+1}, t_{(i-1)\gamma+2}, \ldots, t_\gamma\right)$; in the worst case $\hat{t}_{(i-1)\gamma+1}^{(b)} \neq$

$t_{(i-1)\gamma+1}, \ldots, \hat{t}^{(b)}_{i\gamma} \neq t_{i\gamma}$. This implies that for every (i,ℓ) there are at most γ instances where $\hat{t}^{(b)}_v \neq t_v$, which gives the desired result. ∎

As a result of the previous lemma, it follows that we can recover a binary vector $\hat{\mathbf{t}}^{(b)}$ that is within Hamming distance $\frac{dk}{4}$ of the binary test result $\mathbf{t}$ based on $\mathbf{B}$. Thus, we have to recover the set of infected individuals given a noisy set of test outcomes. To correct errors, we make use of the test outcome generated by the matrix $\mathbf{S}^{(2)}$; this matrix renders the errors in $\mathbf{t}$ "asymmetric," which simplifies the problem. Here, the term "asymmetric" refers to the fact to be addressed in Claim 7 that $\hat{\boldsymbol{t}} \geqslant \boldsymbol{t}$ so that in $\boldsymbol{t}$ a 0 can change to a 1 but not otherwise. More precisely, we use $\mathbf{S}^{(2)}$ to identify tests in $\mathbf{S}^{(1)}$ that contain $> 4\gamma$ defectives. Note that if at least 4γ infected individuals are present in some test pool i, then the entries indexed by $(i-1)\gamma+1, (i-1)\gamma+2, \ldots, i\gamma$ of $\hat{\mathbf{t}}^{(b)}$ may be in error.

Let $\mathbf{s}^{(2)} = (s^{(2)}_1, \ldots, s^{(2)}_{m/\gamma}) \in [0, \tau-1]^{\frac{m}{\gamma}}$ be the test outcomes of $\mathbf{S}^{(2)}$. Define a vector

$$\bar{t}^{(b)}_j = \begin{cases} 1, & \text{if } s^{(2)}_{\lceil \frac{j}{m} \rceil} \geqslant 4\gamma, \\ \hat{t}^{(b)}_j, & \text{otherwise.} \end{cases} \tag{4}$$

Similarly as before, for $\mathbf{s} = (\mathbf{s}^{(1)}, \mathbf{s}^{(2)})$ we write

$$\overline{f}_{\tau \to b}(\mathbf{s}) = \bar{\mathbf{t}}^{(b)}.$$

The following straightforward claim follows from the previous discussion and the observations in Claim 6.

Claim 7. *Let $\bar{\mathbf{t}}^{(b)} = \overline{f}_{\tau\to b}(\mathbf{s})$. Then, $\bar{\mathbf{t}}^{(b)} \geqslant \mathbf{t}$, and*

$$d_H(\bar{\mathbf{t}}^{(b)}, \mathbf{t}) \leqslant \frac{dk}{4}.$$

We next generate a list L of potentially infected individuals consistent with the outcome of the tests $\bar{\mathbf{t}}^{(b)}$. The next lemma, which uses the same ideas as Lemma 1, describes an upper bound on the size of L.

Lemma 8. *Suppose that $\mathbf{s} \in [0, \tau-1]^{2\frac{m}{\gamma}}$ is the result of the tests in (2) and (3) and $\bar{\mathbf{t}}^{(b)} = \overline{f}_{\tau\to b}(\mathbf{s})$. Then the size of any list of defectives from $\mathcal{P}$ consistent with $\bar{\mathbf{t}}^{(b)} = f_{\tau\to b}(\boldsymbol{s})$ is at most $\mathcal{O}(d)$.*

Proof: Recall that in our setup the graph $\mathcal{G}$, which is used to construct $\mathbf{B}$ and also $\mathbf{S}^{(1)}$, is an (α,β)-expander. Hence, every vertex in $\mathcal{P}$ has k neighbors and $\beta > \frac{3k}{4}$. As before, let $I \subseteq \mathcal{P}$ denote the set of infected individuals such that $|I| \leqslant d$. Let $\mathbf{t}_I \in \{0,1\}^m$ be the output of the tests dictated by $\mathbf{B}$. We show that given a $S \subseteq \mathcal{P}$ such that $S \cap I = \varnothing$ and $|S| \geqslant \mathcal{O}(d)$, $S \cup I$ cannot be consistent with $\bar{\mathbf{t}}^{(b)}$ under $\mathbf{B}$.

Let $S' = S \cup I \subseteq \mathcal{P}$. Using the same arguments as in the proof of Lemma 1, we can show that the number of unique neighbors of S' satisfies

$$N_u(S') \geqslant \frac{k|S|}{2}.$$

Let $E = \{j : \bar{t}^{(b)}_j > t_j\}$. Since $N(I) \leqslant dk$ and $|E| \leqslant \frac{dk}{4}$ from Claim 7, it follows that

$$\left| N_u(S') \backslash (I \cup E) \right| \geqslant \frac{k|S|}{2} - (dk + \frac{dk}{4}),$$

which implies that if $|S| > \frac{10d}{8}$, then there exists a unique neighbor of S' which is not in error and is also not already covered by an element in I. This implies that $N(S')$ is not consistent with $\bar{\mathbf{t}}^{(b)}$. ∎

The following theorem follows from the previous discussion and from Claim 6 and Lemma 8.

Theorem 9. *There exists a nonbinary two-stage GT scheme that given $\tau = (4\gamma)^\gamma$ thresholds and*

$$\mathcal{O}\Big(\frac{8e^2 d}{\gamma} \log \frac{n}{d}\Big)$$

tests that can identify a set of infected individuals of size at most d in a population of size n.

Proof: We prove the result by describing a simple method for recovering a set L of size $\mathcal{O}(d)$ which contains the set of defectives I. First, we generate the vector $\bar{\mathbf{t}}^{(b)} = \overline{f}_{\tau\to b}(\mathbf{s})$ from our non-binary test outcomes. We initialize $L = \varnothing$. Then, for every $p \in \mathcal{P}$, if $L \cup \mathcal{P}$ is consistent with $\bar{\mathbf{t}}^{(b)}$, we update $L = L \cup \{p\}$. Otherwise, we do not change L. At the end of this process we have $I \subseteq L$. Furthermore, according to Lemma 8, $|L| \leqslant \mathcal{O}(d)$. The result now follows from Theorem 5. ∎

IV. Nonadaptive SQGT

We describe next constructive nonadaptive testing schemes, which in the asymptotic regime require at most $\mathcal{O}(\frac{d^2}{\gamma}\log n)$ tests, with $\tau = (4\gamma)^\gamma$. Our approach builds upon the construction by Porat and Rothschild (PR construction) [18], which makes use of non-binary error-correcting codes. Our key result is described in Lemma 13.

Let $\mathcal{C} \in \mathbb{F}_q^{m/q}$ be a q-ary linear error-correcting code, where q is an odd prime, of minimum distance $\delta\frac{m}{q}$, $q = \mathcal{O}(d)$, and dimension $\log_q(n)$. The PR construction works by uniquely associating each individual in the population of size n with a codeword in $\mathcal{C}$. Under this setup, the test matrix $\mathbf{B}^{(PR)} = (b_{(c,x),j})_{c\in[1,\frac{m}{q}],x\in[0,q-1],j\in[1,n]}$ is defined as

$$b_{(c,x),j} = \begin{cases} 1, \text{ if } \mathbf{x}^{(j)}_c = x, \\ 0, \text{ otherwise,} \end{cases}$$

where $\mathbf{x}^{(j)}$ is the j-th codeword of $\mathcal{C}$. In words, the test indexed by (c,v) contains the codewords (individuals) from $\mathcal{C}$ whose c-th coordinate equals x.

Our approach for designing a nonadaptive testing scheme is similar to that for the adaptive setting. Each test can be generated by taking a linear combination of γ rows of $\mathbf{B}^{(PR)}$. The total number of tests equals $2\frac{m}{q}\lceil\frac{q}{\gamma}\rceil = \mathcal{O}(\frac{m}{\gamma})$, and once again the tests are represented by $\mathbf{S} = \begin{bmatrix}\mathbf{S}^{(1)} \\ \mathbf{S}^{(2)}\end{bmatrix}$, where $\mathbf{S}^{(2)} =$

$\left(s^{(2)}_{(c,r),j}\right)_{c\in\frac{m}{q},r\in[0,\lfloor\frac{q}{\gamma}\rfloor-1],j\in[1,n]}$, is defined as follows:

$$s^{(2)}_{(c,r),j} = \begin{cases} 1, \text{ if } \mathbf{x}^{(j)}_c = [(r-1)\gamma, \min\{r\gamma-1, q-1\}], \\ 0, \text{ otherwise.} \end{cases}$$

In words, the test in $\mathbf{S}^{(2)}$ indexed by (c,r) contains the codewords (individuals) from $\mathcal{C}$ whose c-th coordinate has a value between $(r-1)\gamma$ and the minimum of $r\gamma-1, q-1$. Note that the reason for using the minimum in the previous range of values is a consequence of the fact that we assumed q to be an odd prime. The tests in $\mathbf{S}^{(1)}$ are defined similarly: Suppose that $x^{(j)}_c = (r-1)\gamma + v'$ where $v' \in [0, \gamma-1]$. Then,

$$s^{(1)}_{(c,r),j} = (4\gamma)^{v'}.$$

For shorthand, we refer to the codewords in the (c,r)-th test in $\mathbf{S}^{(1)}$ as $T_{(c,r)} \subseteq \mathcal{C}$.

Claim 10. *Suppose that the number of infected individuals in the test indexed by (c,r) is at most $4\gamma-1$ so that*

$$\left|T_{(c,r)} \cap I\right| \leqslant 4\gamma - 1.$$

Then, given the output of the test $T_{(c,r)}$ we can uniquely determine

$$\left|\{\boldsymbol{x} \in I : x_c = x\}\right|,$$

for $x \in [(r-1)\gamma, r\gamma-1]$.

Let $\mathbf{1^n}$ denote the all-ones vector of length n. We assume that our code $\mathcal{C}$ is such that $\mathbf{1^n} \in \mathcal{C}$. Henceforth, let

$$\overline{I} = \left\{\mathbf{y} + i \cdot \mathbf{1}^n : \mathbf{y} \in I\right\} \tag{5}$$

for all $i \in \{-\gamma+1, \ldots, -1, 0, 1, \ldots, \gamma-1\}$.

Claim 11. *Let $\mathbf{x} \in \mathcal{C}\backslash I$ be such that $\mathbf{x} \in T_{(c,r)}$. Suppose that for an integer ℓ we have*

$$\left|\{\mathbf{z} \in \overline{I} : x_c = z_c\}\right| \leqslant \ell.$$

Then,

$$\left|T_{(c,r)} \cap I\right| \leqslant \ell.$$

Proof: This follows since if $\mathbf{x} \in T_{(c,r)}$, then $x_c = r(\gamma-1) + v'$ for some $v' \in [0, \gamma-1]$. If $\mathbf{z} \in T_{(c,r)} \cap I$, then $z_c = r(\gamma-1) + v''$ for some $v'' \in [0, \gamma-1]$. Since $v', v'' \in [0, \gamma-1]$, it follows that $z_c + (v'-v'') = r(\gamma-1) + v' = x_c$ where $(v'-v'') \in \{-\gamma+1, \ldots, -1, 0, 1, \ldots, \gamma-1\}$. This in turn implies that $z_c + (v'-v'')$ is the value of component c of a vector from the set $\overline{I}$. ∎

We also need the following result.

Claim 12. *Suppose that $\mathbf{x} \in \mathcal{C}\backslash I$ is such that $\mathbf{x} \in T_{(c,r)}$. If there exists an index $c \in [n]$ satisfying*

$$\left|\{\mathbf{z} \in \overline{I} : x_c = z_c\}\right| \leqslant 4\gamma - 1, \tag{6}$$

and

$$\left|\{\mathbf{z} \in I : x_c = z_c\}\right| = 0, \tag{7}$$

then given the output of the tests dictated by $\mathbf{S}^{(1)}$,$\mathbf{S}^{(2)}$ we can determine that $\mathbf{x} \notin I$.

Proof: From Claim 11 and if (6) holds, we have that $\left|\{I \cap T_{(c,r)}\}\right| \leqslant 4\gamma - 1$. Then from Claim 10, since the number of infected individuals in $T_{(c,r)}$ is at most $4\gamma-1$, we have $\left|\{\mathbf{z} \in I : z_c = x_c\}\right| = 0$ using the test outputs of $T_{(c,r)}$. ∎

Lemma 13. *If $\mathcal{C}$ has minimum distance $\delta > 1 - \frac{1}{2d}$, the tests $\mathbf{S}$ uniquely determine the set I of defectives.*

Proof: According to Claim 12, we need to show that (6) and (7) hold for any $\mathbf{x} \in \mathcal{C}\backslash I$. We start by showing that (7) holds. In particular, we show a stronger claim that there exists a set $C^{(1)} \subseteq \frac{m}{q}$ of size at least $\frac{m}{2q}+1$ where for any $c \in C^{(1)}$, we have

$$x_c \neq y_c, \tag{8}$$

where $\mathbf{y} = (y_1, \ldots, y_{\frac{m}{q}}) \in I$. Note that this implies that the number of coordinates of $\mathbf{x}$ which agree in value with an element of I is at most $\frac{m/q}{2} - 1$. Since any two elements in $\mathcal{C}$ can agree in at most $(1-\delta)\frac{m}{q}$ coordinates and $\delta > 1 - \frac{1}{2d}$, it follows that

$$\left|\{c : x_c = y_c, \mathbf{y} \in I\}\right| \leqslant d(1-\delta)\frac{m}{q} < \frac{m}{2q}.$$

Next, we show that for at least one coordinate in $C^{(1)}$, (6) holds as well. First, note that

$$\left|\left\{(\mathbf{y}, c) \in \overline{I} \times \frac{m}{q} : x_c = y_c\right\}\right| \leqslant 2\gamma d(1-\delta)\frac{m}{q} < \gamma\frac{m}{q},$$

so that for a randomly chosen coordinate $c \in \frac{m}{q}$,

$$E\left[\left|\{\mathbf{y} \in \overline{I} : x_c = y_c\}\right|\right] < \gamma.$$

Invoking Markov's inequality we get

$$\Pr\left(\left|\{\mathbf{y} \in \overline{I} : x_c = y_c\}\right| \geqslant 4\gamma\right) < \frac{1}{4}.$$

Therefore, it follows that there exists a set of coordinates $C^{(2)} \subseteq \frac{m}{q}$ of size at least $\frac{m}{2q}$ such that for any $c \in C^{(2)}$

$$\left|\left\{\mathbf{y} \in \overline{I} : x_c = y_c\right\}\right| < 4\gamma.$$

Since $|C^{(2)}| \geqslant \frac{m}{2q}$ and $C^{(1)} \geqslant \frac{m}{2q}+1$, it follows that $|C^{(1)} \cap C^{(2)}| \geqslant 1$. Letting $c^* \in C^{(1)} \cap C^{(2)}$ we have $\left|\{\mathbf{y} \in \overline{I} : x_{c^*} = y_{c^*}\right| \leqslant 4\gamma - 1$ and $\left|\{\mathbf{y} \in I : x_{c^*} = y_{c^*}\}\right| = 0$. By Claim 12, we conclude that $\mathbf{x} \notin I$. ∎

Open Problems. Despite only a small gap remaining between the lower bound and the actual constructions for the saturation model, many other problems remain open and include:

- Extending the nonadaptive and two-round constructions for general quantization thresholds under the SQGT model;
- Deriving bounds and test strategies for consecutive defective models [26], [27], as these capture the order of arrivals into testing queues;
- Addressing generalized binomial SQGT algorithms [28].

REFERENCES

[1] R. Dorfman, "The detection of defective members of large populations," *Annals of Mathematical Statistics*, vol. 14, pp. 436–440, 1943.

[2] W. Kautz and R. Singleton, "Nonrandom binary superimposed codes," *IEEE Transactions on Information Theory*, vol. 10, pp. 363–377, 1964.

[3] D.-Z. Du and F.-K. Hwang, *Pooling Designs and Nonadaptive Group Testing*. World Scientific, 2006.

[4] H. A. Inan, P. Kairouz, M. Wootters, and A. Özgür, "On the optimality of the Kautz-Singleton construction in probabilistic group testing," *IEEE Transactions on Information Theory*, vol. 65, no. 9, pp. 5592–5603, 2019.

[5] J. Wolf, "Born again group testing: Multiaccess communications," *IEEE Transactions on Information Theory*, vol. 31, no. 2, pp. 185–191, 1985.

[6] A. Dyachkov, "Lectures on designing screening experiments," 2004, lecture Note Series 10.

[7] P. Damaschke, "Threshold group testing," in *General Theory of Information Transfer and Combinatorics*, ser. Lecture Notes in Computer Science, vol. 4123, 2006, pp. 707–718.

[8] B. Lindstrom, "Determining subsets by unramified experiments," *A Survey of Statistical Design and Linear Models*, 1975.

[9] D.-Z. Du and F. Hwang, *Combinatorial Group Testing and its Applications*, 2nd ed. World Scientific, 2000.

[10] A. Emad, J. Shen, and O. Milenkovic, "Symmetric group testing and superimposed codes," in *2011 IEEE Information Theory Workshop*, 2011, pp. 20–24.

[11] A. Emad and O. Milenkovic, "Semiquantitative group testing," *IEEE Transactions on Information Theory*, vol. 60, no. 8, pp. 4614–4636, 2014.

[12] ——, "Code construction and decoding algorithms for semi-quantitative group testing with nonuniform thresholds," *IEEE Transactions on Information Theory*, vol. 62, no. 4, pp. 1674–1687, 2016.

[13] A. G. D'yachkov and V. V. Rykov, "A coding model for a multiple-access adder channel," *Probl. Perdachi Inform.* , pp. 26–32, 1981, in Russian.

[14] A. Dyachkov and V. Rykov, "A survey of superimposed code theory," *Problems of Control and Information Theory*, vol. 12, no. 4, pp. 229–242, 1983.

[15] R. Gabrys, S. Pattabiraman, V. Rana, J. ao Ribeiro, M. Cheraghchi, V. Guruswami, and O. Milenkovic, "AC-DC: Amplification curve diagnostics for Covid-19 group testing," 2020, arXiv:2011.05223.

[16] F. Hwang, "A generalized binomial group testing problem," *Journal of the American Statistical Association*, vol. 70, no. 352, pp. 923–926, 1975.

[17] H. Q. Ngo, E. Porat, and A. Rudra, "Efficiently decodable error-correcting list disjunct matrices and applications," in *International Colloquium on Automata, Languages, and Programming*. Springer, 2011, pp. 557–568.

[18] E. Porat and A. Rothschild, "Explicit nonadaptive combinatorial group testing schemes," *IEEE Transactions on Information Theory*, vol. 57, no. 12, pp. 7982–7989, 2011.

[19] P. Indyk, H. Q. Ngo, and A. Rudra, "Efficiently decodable non-adaptive group testing," in *Proceedings of the twenty-first annual ACM-SIAM symposium on Discrete Algorithms*. SIAM, 2010, pp. 1126–1142.

[20] M. Cheraghchi, "Noise-resilient group testing: Limitations and constructions," *Discrete Applied Mathematics*, vol. 161, no. 1, pp. 81–95, 2013, preliminary version in Proceedings of the FCT 2009. arXiv manuscript published in 2008.

[21] A. De Bonis, L. Gasieniec, and U. Vaccaro, "Optimal two-stage algorithms for group testing problems," *SIAM Journal on Computing*, vol. 34, no. 5, pp. 1253–1270, 2005.

[22] M. Cheraghchi and V. Nakos, "Combinatorial group testing schemes with near-optimal decoding time," in *Proceedings of the 61st Annual IEEE Symposium on Foundations of Computer Science (FOCS)*, 2020.

[23] Z. Füredi, "On r-cover-free families," *Journal of Combinatorial Theory, Series A*, vol. 73, no. 1, pp. 172–173, 1996.

[24] V. Guruswami, C. Umans, and S. Vadhan, "Unbalanced expanders and randomness extractors from Parvaresh-Vardy codes," *Journal of the ACM*, vol. 56, no. 4, 2009.

[25] M. Capalbo, O. Reingold, S. Vadhan, and A. Wigderson, "Randomness conductors and constant-degree expansion beyond the degree/2 barrier," in *Proceedings of the 34th Annual ACM Symposium on Theory of Computing (STOC)*, 2002, pp. 659–668.

[26] T. V. Bui, M. Cheraghchi, and T. D. Nguyen, "Improved algorithms for non-adaptive group testing with consecutive positives," *arXiv preprint arXiv:2101.11294*, 2021.

[27] C. J. Colbourn, "Group testing for consecutive positives," *Annals of Combinatorics*, vol. 3, no. 1, pp. 37–41, 1999.

[28] F. Hwang, "A generalized binomial group testing problem," *Journal of the American Statistical Association*, vol. 70, no. 352, pp. 923–926, 1975.